

# Posterior Exploration based Sequential Monte Carlo for Global Optimization

Bin Liu

## Abstract

We propose a global optimization algorithm based on the Sequential Monte Carlo (SMC) sampling framework. In this framework, the objective function is normalized to be a probabilistic density function (pdf), based on which a sequence of annealed target pdfs is designed to asymptotically converge on the set of global optima. A sequential importance sampling (SIS) procedure is performed to simulate the resulting targets, and the maxima of the objective function is assessed from the yielded samples. The disturbing issue lies in the design of the importance sampling (IS) pdf, which crucially influences the IS efficiency. We propose an approach to design the IS pdf online by embedding a posterior exploration (PE) procedure into each iteration of the SMC framework. The PE procedure can also explore the important regions of the solution space supported by the target pdf. A byproduct of the PE procedure is an adaptive mechanism to design the annealing temperature schedule online. We compare the performance of the proposed algorithm with those of several existing related alternatives by applying them to over a dozen standard benchmark functions. The result demonstrates the appealing properties of our algorithm.

## Index Terms

Global optimization, Sequential Monte Carlo, adaptive annealing schedule, mixture model, particle filter optimization, student's t distribution, Expectation-Maximization

## I. INTRODUCTION

**S**IMULATION based algorithms consist of a branch of population-based approaches to solve optimization problems. Different from the other branches of population-based methods such as particle swarm optimization [1,2], genetic algorithm [3] and ant colony optimization [4,5], the basic idea of simulation based algorithms is to simulate a particular probabilistic density function (pdf) over the promising areas of the global optimum. In particular, if the objective function  $f$  is non-negative and, consequently, we search for its maximum, we can sample from a pdf  $\pi(\cdot)$ , resulting from the normalization of  $f$ , since the mode of the resulting distribution will capture the maximum point we are searching for.

The degree of difficulty in using simulation based methods to find the global optimum largely depends on the structure of the normalized objective function  $\pi(\cdot)$ . If  $\pi(\cdot)$  is a one dimensional unimodal distribution, it only needs to run dozens of steps of a benchmark Metropolis-Hastings sampling to get a satisfactory answer. If  $\pi(\cdot)$  is a multivariate multimodal distribution, it could become challenging to develop a suitable simulation algorithm to yield a satisfactory answer fast.

B. Liu is with Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, 210023 China e-mail: bins@ieee.org.

Manuscript received Nov. 1, 2015; revised July 10, 2016.

There are mainly two directions to resolve the above challenge. The first direction is to introduce a set of smoothly transitional target distributions, which is expected to asymptotically converge on the set of global optima, and then simulate the target distributions sequentially. The motivation is to track closely the converging sequence of distributions. The celebrated simulated annealing (SA) algorithm just falls into the above framework. In particular, the standard SA applies the Boltzmann distributions, which is featured by a parameter termed annealing temperature, to serve as the transitional target distributions [6, 7]. At each iteration, SA essentially simulates a Markov chain whose stationary distribution is the Boltzmann distribution of the current temperature, and the current state becomes the initial state for a new chain at the next iteration. Recently a Sequential Monte Carlo-Simulated Annealing (SMC-SA) algorithm has been proposed, in which the Markov Chain Monte Carlo (MCMC) procedure embedded in the standard SA algorithm is replaced by an importance sampling (IS) procedure [8]. Different from MCMC, the IS procedure allows for parallel implementations and easy assessment of the Monte Carlo error [9, 10]. On the other hand, the main shortcoming of IS approaches is an acute sensitivity to the choice of the IS pdf combined with the fact that it is impossible to come up with a universally efficient IS density [9, 10]. Further, in all SA based methods, the temperature has to decrease slowly enough such that the target distribution does not vary too much from iteration to iteration, which ensures the overall convergence [11, 12]. However a slowly decreasing temperature schedule indicates more iterations, leading to more computing burdens. In practice the temperature schedule is often specified beforehand by human experts, while it is desirable if an adaptive mechanism can be developed to yield an appropriate temperature schedule online.

The second direction to resolve the above challenge resulted from a complex  $\pi(\cdot)$  is to design a model that explicitly defines a probability distribution  $q$  over candidate solutions. The estimation of distribution algorithms (EDAs) [13], evolutionary strategy (ES) [14–16], cross entropy (CE) methods [17, 18] and the particle filter optimization (PFO) algorithm [19–21] all belong to this branch, and thus we call them model based approaches in what follows. In these approaches, a random data set is first generated and then a model is created based on this random data set. The model is then sampled to generate new candidate solutions. At each iteration, the model is updated in order to increase the model quality, ensuring that it will generate better candidate solutions in the following iteration. In EDAs, probabilistic models such as probability vectors and Bayesian networks are built from samples of high quality solutions [13]. In ES, new candidate solutions are often generated from a single normal distribution centered around the best-so-far candidate solution or from a mixture of normal distributions centered around a population of high-quality solutions that has been found [14]. For both the EDAs and ES, there is no explicit definition of the model quality. Different from them, CE methods define the quality of a model in terms of the Kullback-Leibler (KL) distance. It is a non-symmetric measure of the difference between two probability distributions [17, 18]. Specifically, CE methods try to find a pdf that is closer to  $\pi(\cdot)$  as much as possible and the closeness is measured by the KL distance. The PFO algorithm is proven to be equivalent with CE methods, see details in [19]. The efficiency of such model based approaches largely depends on the choice of the model structure and the model parameter updating mechanism. A labor intensive tuning process is usually required before running such algorithms. Further, they are often trapped into local optimum in the searching process.

Motivated by the observation that both the above two directions have their respective limitations, in this paper, we develop a novel simulation based algorithm, termed posterior exploration based SMC (PE-SMC). This algorithm improves state-of-the-art SMC-SA algorithm [8] by fusing a posterior exploration procedure into the framework of SMC. This PE procedure has three

functions. First, it can update the IS density model adaptively at each iteration of SMC. Second it can explore the promising regions of the candidate solutions in both macroscopic and microscopic scales. Lastly, the output of the PE procedure can be used for selecting the annealing temperature of the following iteration.

The remainder of the paper is organized as follows. In Section II, we give a brief overview on the SMC-SA optimization algorithm [8]. In Section III, we present the proposed PE-SMC algorithm. In Section IV, we give a convergence analysis for the PE-SMC algorithm. In Section V, we evaluate the performance of our algorithm in solving a number of benchmark optimization problems. Finally, we conclude the paper in Section VI.

## II. REVISITING SMC-SA

In this section, we present a brief overview on the SMC-SA algorithm [8]. The purpose is to define the notions and introduce the necessary background information for describing the proposed PE-SMC algorithm in Section III. In this paper, we are concerned with the following maximization problem

$$\max_{x \in \chi} f(x) \quad (1)$$

where  $\chi$  denotes the nonempty solution space defined in  $\mathbb{R}^n$ , and  $f: \chi \rightarrow \mathbb{R}$  is a continuous real-valued function. The basic assumption here is that  $f$  is bounded on  $\chi$ , which means  $\exists f_l > -\infty, f_u < \infty$  such that  $f_l \leq f(x) \leq f_u, \forall x \in \chi$ . We denote the maximal function value as  $f^*$ , i.e., there exists an  $x^*$  such that  $f(x) \leq f^* \triangleq f(x^*), \forall x \in \chi$ .

A sequence of target pdfs,  $\pi_1(f), \pi_2(f), \dots$ , is built as follows

$$\pi_k(x) \triangleq \exp\left(\frac{f(x)}{T_k}\right), k = 1, 2, \dots, \quad (2)$$

where  $T_k$  denotes the cooling temperature at the  $k$ th iteration satisfying  $T_k < T_{k-1}$  if  $k > 1$ . A working flow of the algorithm is summarized as follows.

Algorithm 1: SMC-SA

- Initialization (i.e., at iteration 0): Generate  $x_0^i \stackrel{\text{iid}}{\sim} \text{Unif}(\chi), i = 1, 2, \dots, N_0$ .  $N_0$  and  $N_k$  in what follows denote the sample size and belong to the input of the algorithm. Set  $k = 1$ .
- At iteration  $k$ :
  - Importance weight updating: calculate normalized importance weights according to  $\omega_k^i \propto \pi_1(x_0^i)$  if  $k = 1$ , and  $\omega_k^i \propto \frac{\pi_k(x_{k-1}^i)}{\pi_{k-1}(x_{k-1}^i)}$  if  $k > 1$ .
  - Resampling: draw i.i.d. samples  $\{\tilde{x}_k^i\}_{i=1}^{N_k}$  from a distribution  $\tilde{\pi}_k(x) \triangleq \sum_{i=1}^{N_{k-1}} \omega_k^i \delta(x - x_{k-1}^i)$ , where  $\delta$  denotes the Dirac delta function.
  - One step of Metropolis sampling:
    - \* Choose a symmetric proposal distribution with density  $g_k(\cdot|x)$ , such as a normal distribution with mean  $x$ .
    - \* Generate  $y_k^i \sim g_k(y|\tilde{x}_k^i), i = 1, \dots, N_k$ .
    - \* Calculate acceptance probability

$$\rho_k^i = \min\left\{\frac{\pi_k(y_k^i)}{\pi_k(\tilde{x}_k^i)}, 1\right\}.$$

\* Accept

$$x_k^i = \begin{cases} y_k^i, & \text{with probability } \rho_k^i; \\ \tilde{x}_k^i, & \text{with probability } 1 - \rho_k^i. \end{cases}$$

- Stopping: if a stopping criterion is satisfied, return  $\max_i f(x_k^i)$ ; otherwise, set  $k = k + 1$  and continue.

As is shown above, the SMC-SA algorithm tracks the converging sequence of Boltzmann distributions using a population of samples via SMC method, such that the empirical distributions yielded by SMC-SA also converge weakly to the uniform distribution concentrated on the set of global optima as the temperature decreases to zero.

A convergence analysis for the SMC-SA algorithm is given in [8], which indicates that an appropriate choice of the sample size sequence  $\{N_k\}$  and the temperature cooling schedule  $\{T_k\}$  is required. There is a fundamental trade-off between the temperature change rate and the sample size. Conceptual guidance is provided in [8] for selecting the cooling schedule and the sample size, while it still lacks a specific working method for use in practice.

In addition, it is shown that the SMC-SA algorithm only depends on one step of Metropolis sampling procedure to generate new candidate solutions, while this operation only moves each candidate solution to its local neighborhood in the solution space  $\chi$  with a probability. Such a local searching mechanism is not good enough for finding the global optima, especially when the optimal solution lies far from the neighborhood of the solutions found so far; because the search may stop due to the impossibility of improving the solution, even if the best solution found so far is not optimal. Lastly the Boltzmann distributions may lead to numerical problems due to their exponential form, as pointed out in [8].

### III. THE PROPOSED PE-SMC ALGORITHM FOR GLOBAL OPTIMIZATION

In this section, we present the PE-SMC algorithm in detail. The objective function  $f(x)$  is assumed to be positive for any  $x$ . The series of target pdfs is defined to be

$$\pi_k(x) \propto f(x)^{\lambda_k}, k = 1, 2, \dots, \quad (3)$$

where  $k$  denotes the index of the target pdf,  $\lambda \in \mathbb{R}$  the annealing temperature variable which satisfies  $0 < \lambda_{k-1} < \lambda_k$ . A working flow of our algorithm is summarized as follows.

#### Algorithm 2: PE-SMC

- Initialization (iteration 0): initialize the IS density function  $q$  with parameter value  $\psi_0$ . Denote the resulting pdf by  $q(\cdot|\psi_0)$ . Note that we select here a mixture of Student's  $t$  pdfs to model  $q$ , see details in Sec.III-A. Initialize the sample size  $N$  and the cooling temperature  $\lambda_1$ . Set  $k = 1$ .
- At iteration  $k$ :
  - Draw i.i.d. random samples  $x_k^i, i = 1, 2, \dots, N$  from  $q(\cdot|\psi_{k-1})$ .
  - Perform the PE procedure and output the revised parameter of  $q$ , denoted by  $\psi_k$ . See details in Sec.III-B.
  - Select the cooling temperature for the following iteration, namely  $\lambda_{k+1}$ . See details in Sec.III-C.
  - Stopping: if a stopping criterion is satisfied, return the best solution found so far, otherwise, set  $k = k + 1$  and continue.

In what follows, we introduce each module of the algorithm in detail.

#### A. Initialization

We design the IS density  $q$  to be a mixture of Student's  $t$  pdfs. The use of a mixture type model makes  $q$  easy to account for possible multimodal structures of the target pdfs. The Student's  $t$  pdf is chosen as the mixing components due to its heavy-tail property, which is beneficial for the algorithm to explore the solution space.

Following [22], suppose that  $Y$  is a  $d$  dimensional random variable that follows the multivariate Student's  $t$  distribution, denoted by  $\mathcal{S}(\cdot|\mu, \Sigma, v)$ , where  $\mu$  denotes the mean,  $\Sigma$  a positive definite inner product matrix and  $v \in (0, \infty]$  is the degrees of freedom. Then the density function of  $Y$  is:

$$\mathcal{S}(Y|\mu, \Sigma, v) = \frac{\Gamma(\frac{v+d}{2})|\Sigma|^{-0.5}}{(\pi v)^{0.5d}\Gamma(\frac{v}{2})\{1 + M_d(Y, \mu, \Sigma)/v\}^{0.5(v+d)}}, \quad (4)$$

where

$$M_d(Y, \mu, \Sigma) = (Y - \mu)^T \Sigma^{-1} (Y - \mu) \quad (5)$$

denotes the Mahalanobis squared distance from  $Y$  to  $\mu$  with respect to  $\Sigma$ .

As a mixture of Student's  $t$  pdfs,  $q$  is defined to be:

$$q(x|\psi) \triangleq \sum_{m=1}^M \alpha_m \mathcal{S}(x|\mu_m, \Sigma_m, v), \quad (6)$$

where  $m$  denotes the component index,  $M$  the total number of the mixing components,  $\alpha_m$  the probability mass of the  $m$ th component satisfying  $\alpha_m > 0$  for each  $m \in \{1, 2, \dots, M\}$  and  $\sum_{m=1}^M \alpha_m = 1$ , and  $\psi \triangleq \{M, \{\alpha_m, \xi_m\}_{m=1}^M\}$  consists of all the tunable parameters of  $q$ .

The main task of the posterior exploration procedure, which will be presented in detail in the following subsection, is to tailor the shape of  $q$  by updating the value of  $\psi$ , in order to make  $q$  resemble the target distribution to the greatest extent.

#### B. The posterior exploration procedure

The PE procedure is where our algorithm mainly differs with existing alternatives. The purpose of this procedure is to facilitate exploring target distribution structures that have not been represented enough by the IS pdf. As the target distribution is termed the posterior in Bayesian statistics, we call this procedure posterior exploration (PE).

The PE procedure runs four operators in turn, namely IS, componentwise Metropolis move [23], expectation-maximization (EM) estimation and addition of new mixing components. Now we introduce the operators one by one in the following.

1) *The IS operator:* IS is one of the basic Monte Carlo techniques for simulating a target distribution [9, 24]. The input of this operator consists of a IS pdf  $q(\cdot|\psi)$ , a target pdf  $\pi$  and the sample size  $N$ . Given the input, the IS operator first draws a set of random samples,  $x_k^i$ ,  $i = 1, 2, \dots, N$ , from  $q$ , and then weights the samples as follows,

$$\tilde{\omega}^i = \frac{\pi(x^i)}{q(x^i)}, i = 1, 2, \dots, N. \quad (7)$$

A normalization operation is finally performed,

$$\omega^i = \frac{\tilde{\omega}^i}{\sum_{i=1}^N \tilde{\omega}^i}, i = 1, 2, \dots, N. \quad (8)$$

The output of the IS operator is just a weighted sample set  $\{x^i, \omega^i\}_{i=1}^N$ .

2) *the componentwise Metropolis moving operator*: The input of this operator consists of the weighted sample set  $\{x^i, \omega^i\}_{i=1}^N$  and the corresponding target pdf  $\pi$ . The purpose of this operator is to explore the target distribution in a microscopic scale. This operator gives a fine tuning to each sample. Given the  $i$ th sample  $x^i$ , the operator tunes each dimension of it one by one. Suppose that the  $j$ th dimension is under consideration, the operator generates a zero-mean Gaussian distributed random number, and then adds it to the  $j$ th dimension of  $x^i$  to generate a new sample. We denote the new sample by  $x^{i'}$  and calculate the acceptance probability

$$\rho^i = \min \left\{ \frac{\pi(x^{i'})}{\pi(x^i)}, 1 \right\}.$$

Finally we accept

$$x^i = \begin{cases} x^{i'}, & \text{with probability } \rho^i; \\ x^i, & \text{with probability } 1 - \rho^i. \end{cases}$$

3) *the EM operator*: The EM mechanism is an effective approach to improve the IS efficiency by tailoring a mixture formed IS density function [10]. Inspired by the success of the application of the EM operator in the adaptive IS algorithm [10], we adopt it here to improve the IS efficiency of the PE-SMC algorithm. The input of the EM operator consists of a weighted sample set  $\{x^i, \omega^i\}_{i=1}^N$ , the corresponding IS pdf  $q(\cdot|\psi)$  and the target pdf  $\pi$ . The EM operator includes an Expectation and an Maximization step as follows [10]:

the Expectation step:

$$\alpha_m' = \sum_{n=1}^N w^i \epsilon(m|x^i). \quad (9)$$

the Maximization step:

$$\mu_m' = \frac{\sum_{i=1}^N w^i \epsilon(m|x^i) u_m(x^i) x^i}{\sum_{i=1}^N \omega^i \epsilon(m|x^i) u_m(x^i)}, \quad (10)$$

$$\Sigma_m' = \frac{\sum_{i=1}^N \omega^i \epsilon(m|x^i) u_m(x^i) C_i}{\sum_{i=1}^N \omega^i \epsilon(m|x^i)}, \quad (11)$$

where

$$u_m(x) = \frac{v + d}{v + (x - \mu_m)^T (\Sigma_m)^{-1} (x - \mu_m)}, \quad (12)$$

$$\epsilon(m|x) = \frac{\alpha_m \mathcal{S}(x|\mu_m, \Sigma_m, v)}{q(x|\psi)}, \quad (13)$$

and  $C_i = (x^i - \mu_m)(x^i - \mu_m)^T$ ,  $d$  denotes the dimension of  $x$ , and  $A'$  denotes the updated value of  $A$ . Note that  $\epsilon(m|x)$  in

Eqn. (13) denotes the probability mass in the event that  $x$  belongs to component  $m$ , and  $\alpha_m'$  in Eqn. (9) is just a weighted summation of those probability masses. Finally, update the parameters  $\alpha_m = \alpha_m'$ ,  $\mu_m = \mu_m'$ ,  $\Sigma_m = \Sigma_m'$ ,  $m = 1, \dots, M$ . More details about the EM operation can be found for example in [10, 25, 26].

4) *Operator for adding new mixing components*: The purpose of this operator is to explore the promising solution space that has not been represented well enough or even totally neglected by the current IS pdf so far. This operator was originally developed by us in [22]. The basic idea is illustrated conceptually in Fig. 1. For simplicity and clarity in presenting the idea, a one dimensional case is considered.

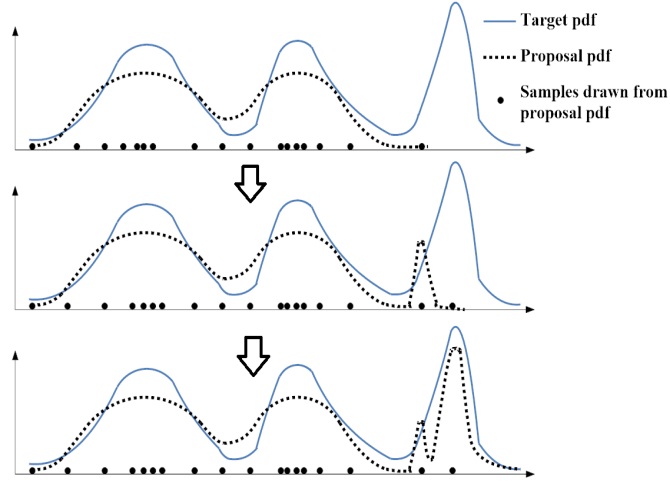


Fig. 1: A conceptual illustration on the operator of adding new mixing components in the IS pdf

The target and the IS pdfs are represented by the solid and dotted curves, respectively. The samples drawn from the IS pdf are represented by solid circles. As is shown in the top figure, at the beginning, the IS pdf does not represent the target structure well enough, since the rightmost peaky mode is not accounted. The above information is revealed by the highest weight sample. As seen in the top figure, the highest weight sample is the rightmost one. The idea is to add a new component centered at the highest weight sample into the mixture IS pdf and then draw a set of samples from the new component, as illustrated in the middle figure. The new samples that come from the new mixing component will further be used to explore the target. Again a new component is added and centered at the highest weight sample. Finally we obtain the bottom plot in Fig.1. It is shown that the operator of adding new mixing component can bring in a kind of ripple effect, which facilitates the algorithm to explore more promising areas in the solution space and finally an IS pdf that can represent the whole structure of the target will be obtained.

In practical implementation, after every ten new components being added, the total parameters of the mixture IS pdf are updated by running the IS and EM operators in turn, as described in Subsections III-B1 and III-B3, respectively. The components that have negligible probability masses will be deleted and correspondingly probability masses of the survival components are increased proportionally with respect to their original values under the constraint that their summation is 1.

The stopping criteria of adding new components is the effective sample size (ESS), which is an empirical measure to evaluate

the IS efficiency [27]. Given a weighted sample set  $\{x^i, \omega^i\}_{i=1}^N$ , the ESS is defined to be

$$\text{ESS} \triangleq \frac{1}{\sum_{i=1}^N (\omega^i)^2}. \quad (14)$$

and the normalized ESS (NESS) is  $\text{ESS}/N$ , which satisfies  $1 \leq \text{NESS} \leq 1$ . The limiting case  $\text{NESS} = 1$  happens only if the IS pdf is designed to be the same as the target and thus all the samples own the same weight  $1/N$ . So we use the NESS to measure the quality of the IS pdf based on the fact that the bigger the NESS the greater extent to which the IS pdf resembles the target. In practice, we check whether the NESS after adding each new mixing component is bigger than a prescribed threshold, if so, stop the process of adding new components, otherwise continue the process.

### C. Adaptive determination of the cooling temperature

Here we present an adaptive approach to determine  $\lambda_{k+1}$  at the end of the  $k$ th iteration of the PE-SMC algorithm. Observe that the annealed target pdf  $\pi(\cdot)$  is a function of  $\lambda$  as shown in Equation (3). Given the IS pdf  $q(\cdot|\psi_k)$  and the random samples drawn from it, the  $k+1$ th target,  $\pi_{k+1}(\cdot)$ , actually determines the importance weights, which then determines an ESS value. Thus the ESS can be treated as a function of  $\lambda$  as follows

$$\text{ESS} \triangleq h(\lambda). \quad (15)$$

Here, based on  $\lambda_k$ , we specify  $\lambda_{k+1}$  as follows

$$\lambda_{k+1} = \arg \min_{\lambda \in (\lambda_k, \infty)} \|h(\lambda) - \beta \cdot \text{ESS}_k\|^2, \quad (16)$$

where  $\text{ESS}_k$  is calculated based on importance weights that are associated with the target pdf  $\pi_k(\cdot)$  and the IS pdf  $q(\cdot|\psi_k)$ , and the parameter  $\beta \in (0, 1)$  is used to control the extent to which  $\pi_{k+1}(\cdot)$  differs from  $\pi_k(\cdot)$ . Empirically we set  $\beta$  to 0.8 to guarantee a smooth transition from  $\pi_k(\cdot)$  to  $\pi_{k+1}(\cdot)$ .

## IV. CONVERGENCE ANALYSIS

In this section, we make a detailed convergence analysis for the PE-SMC algorithm. To begin with, we show that under our assumptions on  $\chi$  and  $f$ , the target distribution converges weakly to the uniform distribution on the set of optimal solutions as the annealing temperature variable  $\lambda$  in Eqn.(3) increases to infinity. In particular, if there is only one unique optimum solution, it converges weakly to a degenerate distribution concentrated on that solution.

Set the *target distribution*  $r$  to be the absolutely continuous probability measure with density  $\pi$ , that is

$$r(A) = \int_A \pi(x) dx, \forall A \in \mathfrak{B} \quad (17)$$

where  $\mathfrak{B}$  denotes the Borel  $\sigma$ -field on  $\chi$ . For  $\varepsilon > 0$ , define the  $\varepsilon$ -level set as follows

$$\chi_\varepsilon = \{x \in \chi : f(x) \geq f^* - \varepsilon\} \quad (18)$$

where  $f^*$  is, as before, the global maximum of  $f$  on  $\chi$ . Then we have



**PROPOSITION 1:** For all  $\varepsilon > 0$ ,  $\lim_{\lambda \rightarrow \infty} r_\lambda(\chi_\varepsilon) = 1$ .

*Proof.* Fix  $\varepsilon$ . Then

$$r_\lambda(\chi_\varepsilon) = 1 - r_\lambda(\chi - \chi_\varepsilon) = 1 - \frac{\int_{\chi - \chi_\varepsilon} f(x)^\lambda dx}{\int_\chi f(z)^\lambda dz} \quad (19)$$

$$\geq 1 - \frac{\int_{\chi - \chi_\varepsilon} f(x)^\lambda dx}{\int_{\chi_{\varepsilon/2}} f(z)^\lambda dz} \quad (20)$$

$$\geq 1 - \frac{(f^* - \varepsilon)^\lambda \phi(\chi - \chi_\varepsilon)}{(f^* - \varepsilon/2)^\lambda \phi(\chi_{\varepsilon/2})} \quad (21)$$

$$= 1 - \left( \frac{f^* - \varepsilon}{f^* - \varepsilon/2} \right)^\lambda \frac{\phi(\chi - \chi_\varepsilon)}{\phi(\chi_{\varepsilon/2})}. \quad (22)$$

where  $\phi$  denotes the Lebesgue measure on  $\mathbb{R}^n$ . (Note that our assumptions on  $\chi$  and  $f$  guarantee that  $\phi(\chi_{\varepsilon/2}) > 0$ .) Thus

$$\lim_{\lambda \rightarrow \infty} r_\lambda(\chi_\varepsilon) = 1.$$

Proposition 1 suggests that, as  $\lambda$  tends to infinity,  $r_\lambda$  converges weakly to a distribution that concentrates on the set

$$\chi_\varepsilon = \{x \in \chi : f(x) \geq f^* - \varepsilon\}.$$

Now fix  $\lambda > 0$  and hence the target pdf  $\pi_\lambda$ . Let's analyze convergence properties of the proposed four operators that constitute the PE procedure. The IS principle suggests that the weighted samples yielded by the IS operator can give a Monte Carlo approximation to the current target pdf  $\pi_\lambda$  [9, 12, 24]. As its target pdf is also  $\pi_\lambda$ , the follow-up componentwise Metropolis moving operator will bring the distribution of the weighted samples closer to  $\pi_\lambda$  and thus combat the approximation error introduced by the IS operator. The IS principle also tells us that the efficiency of the IS operator depends on the choice of the IS density function  $q(\cdot|\varphi)$ . The EM mechanism has proven to be an effective way to improve the quality of an mixture formed IS density function [10]. The main shortcoming of EM lies in that it doesn't allow the number of mixing components  $M$  to be changed, which indicates that, if the value of  $M$  is not selected appropriately beforehand, the EM operator will not find the optimal parameter value for the IS density function. The last operator, namely the new components addition operator, remedies the defect of the EM operator by providing a mechanism to tailor  $M$  online. As this operator increases the IS pdf around the local area centered at the current maximum weight sample, it always increases the ESS value to some extent. Since ESS is the most recognized metric to measure the IS efficiency in SMC methods [11, 12, 28], the validity of the operation of adding new components is guaranteed.

To summarize, at each iteration of PE-SMC, the IS operator gives an initial Monte Carlo approximation to  $\pi_\lambda$ , then the follow-up operators contribute to decreasing the approximation error based on different mechanisms.

Since the random samples simulated with PE-SMC approximate  $\pi_\lambda$  at each iteration and we have already proved that  $\pi_\lambda$  converges weakly to the point mass at the global optimum  $x^*$ , we can conclude that the random samples simulated with PE-SMC will finally converge to  $x^*$  as well.

Test Problems	Goal: $f(x^*)$	PSO	PF	SMC-SA	PE-SMC
TF1	2D 30	<b>30</b> $\pm 0$	29.4447 $\pm 0.53$	29.9943 $\pm 6 \times 10^{-3}$	29.9989 $\pm 6.14 \times 10^{-4}$
TF2	2D 1.56261	<b>1.5626</b> $\pm 2.23 \times 10^{-15}$	1.5625 $\pm 6.14 \times 10^{-5}$	1.5626 $\pm 7.80 \times 10^{-6}$	<i>1.5626</i> $\pm 5.01 \times 10^{-7}$
TF3	2D 1	<i>0.9974</i> $\pm 1.26 \times 10^{-2}$	0.9836 $\pm 1.60 \times 10^{-2}$	<b>0.9978</b> $\pm 8.60 \times 10^{-3}$	0.9960 $\pm 5.50 \times 10^{-3}$
TF4	2D 2459.6407	2410.6 $\pm 75.08$	2429.9 $\pm 26.6326$	<b>2436.9</b> $\pm 24.1397$	<i>2431.4</i> $\pm 30.2447$
TF5	2D 1000	<b>999.9979</b> $\pm 3.30 \times 10^{-3}$	999.8415 $\pm 0.16$	999.9058 $\pm 8.16 \times 10^{-2}$	999.9895 $\pm 8.3 \times 10^{-3}$
TF6	2D 19.2085	<b>19.2085</b> $\pm 4.04 \times 10^{-14}$	19.1981 $\pm 1.33 \times 10^{-2}$	19.2084 $\pm 1.42 \times 10^{-4}$	<i>19.2085</i> $\pm 6.11 \times 10^{-6}$
TF7	2D 100	<b>100</b> $\pm 0$	99.9996 $\pm 5.64 \times 10^{-4}$	<i>100</i> $\pm 4.95 \times 10^{-5}$	<b>100</b> $\pm 0$
TF8	2D 450	<b>450</b> $\pm 0$	449.9759 $\pm 2.92 \times 10^{-2}$	<i>449.9990</i> $\pm 1.20 \times 10^{-3}$	<b>450</b> $\pm 0$
TF9	2D 200	<b>200</b> $\pm 0$	199.9651 $\pm 3.60 \times 10^{-2}$	199.9977 $\pm 3.00 \times 10^{-3}$	<i>199.9999</i> $\pm 1.12 \times 10^{-6}$
	5D 200	<b>200</b> $\pm 0$	197.0448 $\pm 0.62$	199.9854 $\pm 3.67 \times 10^{-2}$	<i>199.9997</i> $\pm 4.38 \times 10^{-4}$
	10D 200	<i>199.8007</i> $\pm 0.4690$	193.9105 $\pm 34.27$	199.6702 $\pm 4.10 \times 10^{-2}$	<b>199.9487</b> $\pm 6.48 \times 10^{-3}$
	20D 200	<i>199.7810</i> $\pm 0.5015$	188.6582 $\pm 44.83$	196.3677 $\pm 5.33 \times 10^{-1}$	<b>199.9228</b> $\pm 1.13 \times 10^{-2}$
TF10	2D 1	<b>1</b> $\pm 0$	<i>1</i> $\pm 3.25 \times 10^{-5}$	1 $\pm 6.55 \times 10^{-5}$	1 $\pm 3.77 \times 10^{-5}$
TF11	2D 1800	1758.5 $\pm 59.22$	1777.6 $\pm 32.09$	<i>1800</i> $\pm 3.19 \times 10^{-4}$	<b>1800</b> $\pm 1.35 \times 10^{-4}$
TF12	2D 486.7309	<b>486.7309</b> $\pm 5.98 \times 10^{-13}$	485.599 $\pm 1.33$	486.7309 $\pm 9.91 \times 10^{-6}$	486.7298 $\pm 9.50 \times 10^{-3}$
TF13	2D 120	<b>120</b> $\pm 0$	119.9994 $\pm 5.16 \times 10^{-4}$	119.9999 $\pm 1.63 \times 10^{-4}$	<i>120</i> $\pm 9.66 \times 10^{-6}$
TF14	2D $1.8 \times 10^5$	<i>1.8 \times 10^5</i> $\pm 2.32 \times 10^{-11}$	$1.8 \times 10^5 \pm 1.37 \times 10^{-2}$	$1.8 \times 10^5 \pm 8.4 \times 10^{-3}$	<b>1.8</b> $\times 10^5 \pm 0$
TF15	2D $\approx 509[31]$	508.9427 $\pm 0.50$	508.9830 $\pm 0.11$	<i>509.0020</i> $\pm 1.93 \times 10^{-10}$	<b>509.0020</b> $\pm 1.50 \times 10^{-11}$
TF16	2D 1	0.7900 $\pm 0.4094$	0.9974 $\pm 5.30 \times 10^{-3}$	0.5000 $\pm 0.50$	<b>1</b> $\pm 1.21 \times 10^{-7}$
TF17	2D 1.8013	<b>1.8013</b> $\pm 2.89 \times 10^{-15}$	1.7982 $\pm 3.10 \times 10^{-3}$	1.8008 $\pm 7.04 \times 10^{-4}$	<i>1.8013</i> $\pm 1.67 \times 10^{-8}$
	5D 4.687658	<i>4.6624</i> $\pm 4.01 \times 10^{-2}$	4.2103 $\pm 0.1648$	4.6267 $\pm 0.1616$	<b>4.6875</b> $\pm 2.81 \times 10^{-4}$
	10D 9.66015	<i>9.4562</i> $\pm 0.16$	5.9357 $\pm 0.2852$	9.3155 $\pm 0.8074$	<b>9.6596</b> $\pm 1.77 \times 10^{-2}$

TABLE I: Convergence results yielded from 100 independent runs of each algorithm on each test problem

## V. PERFORMANCE EVALUATION

### A. Performance comparison with other alternatives

We applied the proposed algorithm, as well as the Trelea type vectorized PSO [29], the PFO algorithm [19, 20] and the SMC-SA algorithm [8], to search the global maximum of over a dozen standard benchmark functions, termed TF1, TF2, ..., TF17 in what follows. See details about these functions in Sec. VII. For some of the test functions, we considered their two dimensional (2D), 5D, 10D, and 20D implementations in order to investigate the abilities of the algorithms in handling higher dimensional cases. Following [29, 30], the swarm size of PSO was set to 50 in all cases, and the swarm was allowed to perform a maximum number of 10000 iterations. For the other algorithms under consideration, the particle size was set to 500 for 2D cases, 2000 for 5D cases, 5000 for 10D cases and 50000 for 20D cases. The degree of freedom  $v$  of all the Student's  $t$  components is fixed to be 5. The termination condition is that in the last 10 iterations, no better solution is found.

Each algorithm is run 100 times independently for each optimization task and the convergence results are listed in Table I. Boldface and italic in the table indicate the best and the second best result obtained, respectively. A simple statistics to summarize the comparison results is presented in Table II. As is shown, for all 23 cases under consideration, PSO yielded the best result for 12 times, while it yielded the worst one for 3 times. The proposed PE-SMC algorithm did the best for 10 cases and it is more attractive than the others in that it never produced the worst performance.

Algorithms	PSO	PFO	SMC-SA	PE-SMC
#Best	12	0	2	10
#2nd Best	6	2	5	9
#Worst	3	17	2	0

TABLE II: Statistics of the performance comparison results. # denotes the number of events, e.g., #Best denotes the number of the event that the corresponding algorithm yielded the best result.

Observe that PE-SMC beats significantly PFO and SMC-SA, which demonstrates the performance gain obtained due to the application of the proposed PE operation in the SMC framework.

After scrutiny on Table I, one can see that, for all 10D and 20D cases and all but one 5D cases, PE-SMC performs the best, which is an indication of that PE-SMC is more preferable in dealing with higher dimensional optimization problems.

### B. Examination of the PE procedure

For availability of visualization, we present intermediate results of applying the PE-SMC algorithm to search maximum of a 2D Rastrigin function defined in Sec.VII-I, in order to examine the effect of the PE procedure. This test function is plotted in Fig.2. The global maximum is 200 which is located at the origin point.

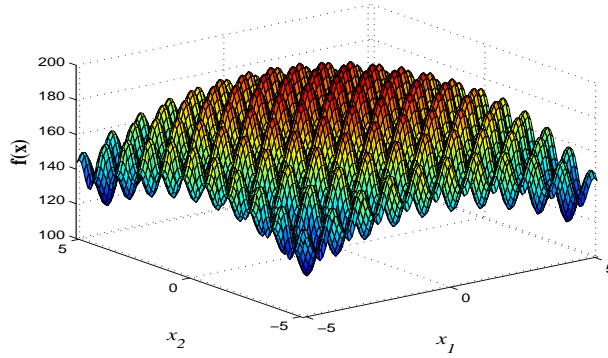


Fig. 2: The 2D Rastrigin function defined in Sec.VII-I

The PE-SMC algorithm is initialized by a one component IS pdf. The IS pdf as well as the samples drawn from it before running the PE procedure at the first iteration is plotted in Fig.3. The contours in the figures represent the location, scale, and shape of the components in the Student's  $t$ -mixture IS pdf, which are drawn at one standard deviation from the means in each of the major and minor axis directions. As is shown, the current IS pdf does not cover well the important regions of the parameter space.

The updated IS pdf and its corresponding samples after running the PE procedure are plotted in Fig.4. As is shown, the updated IS pdf owns more mixing components and its support has already covered the location of the global maximum. Because of the limited sample size and the multimodal structure of the test function as shown in Fig.2, it is still nontrivial to find the global maximum. But as the algorithm iterates, the support of the IS pdf will converge on the local area of the maximum gradually, as shown in Figs.5 and 6, which plot the results after running the PE procedure at the 4th and 7th iterations respectively. Note the scale changes of the figures.

For searching maximum of a 20D Rastrigin function, it is impossible to graphically plot the yielded intermediate IS pdf as before, while we recorded the number of survival mixing components in the IS pdf and now plot it in Fig.7. The increase in the number of mixing components indicates an exploration dominated process, while the decrease corresponds to an exploitation dominated process. Taken together, it is shown that the proposed algorithm brings in a dynamic balance between the exploration and exploitation process. Thanks to this balance, the found optimum value of the test function converges fast to the goal value, 200, as shown in Fig.8.

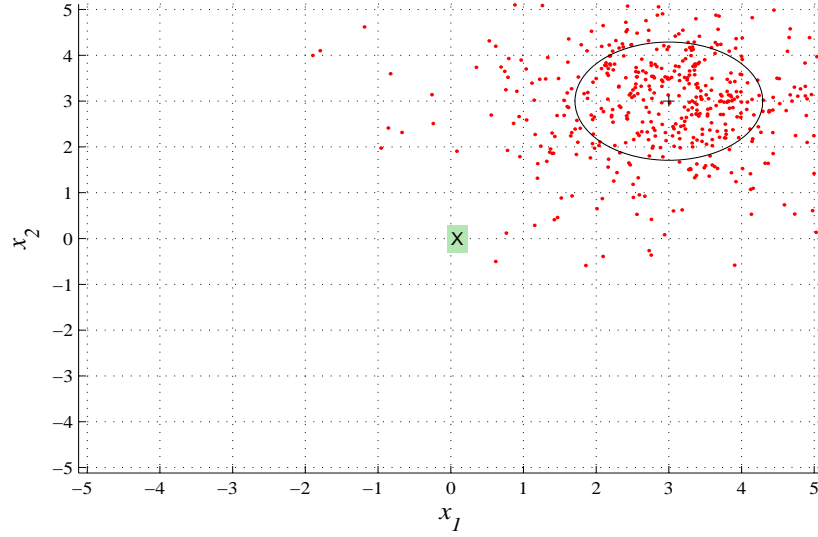


Fig. 3: The IS pdf and the samples drawn from it in the 1st iteration before running the PE procedure

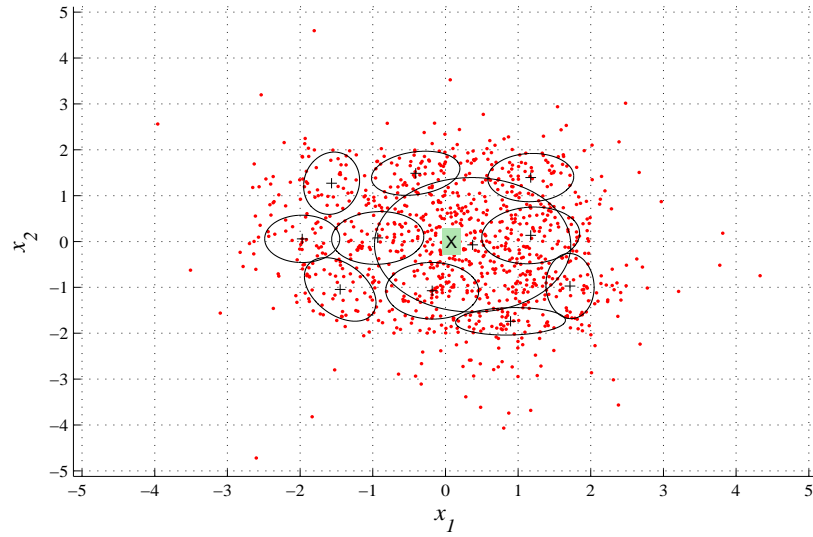


Fig. 4: The IS pdf and samples drawn from it in the 1st iteration after running the PE procedure

### C. Examination on the procedure of adapting the SA temperature schedule

Take the result of an example run of our algorithm in a 20D Rastrigin function optimization task for example, the NESS changes per iteration as graphically shown in Fig.9. Observe that the NESS value fluctuates up and down, while the major trend is to become larger, which indicates that the IS efficiency grows along the adaptation of the SA's temperature. The yielded annealing temperature is plotted in Fig.10. It is shown that the change of the annealing temperature almost conforms with an exponential growth law. We argue that this result is consistent with the theory which states that an exponentially long annealing schedule can guarantee convergence to the global optimum (even for non-convex problems) [32].

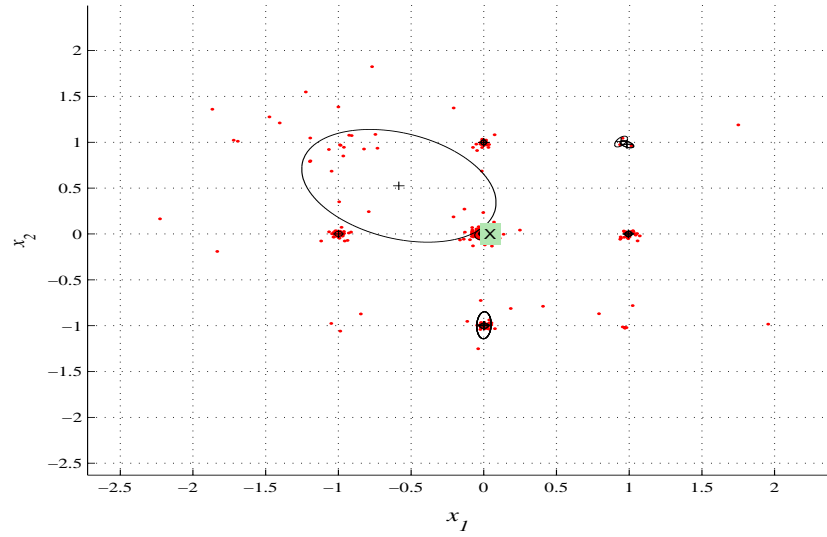


Fig. 5: The IS pdf and samples drawn from it in the 4th iteration after running the PE procedure

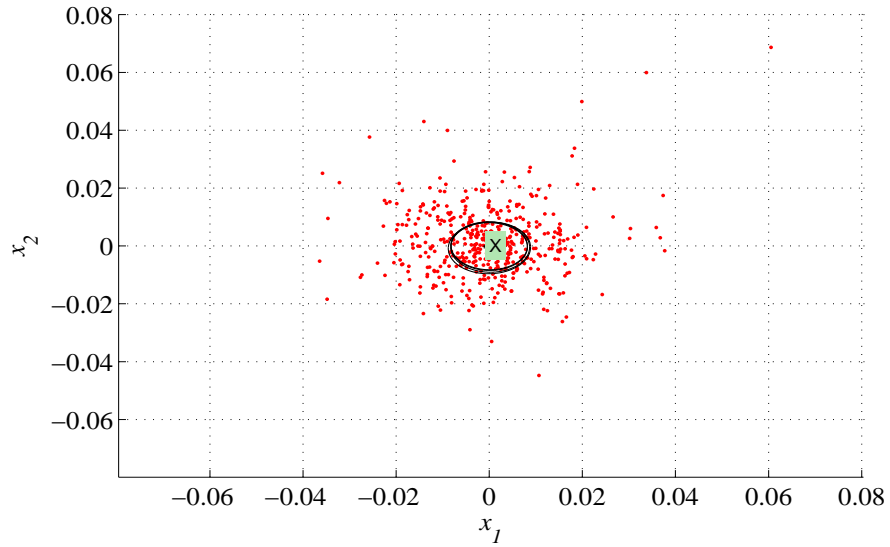


Fig. 6: The IS pdf and samples drawn from it in the last iteration after running the PE procedure

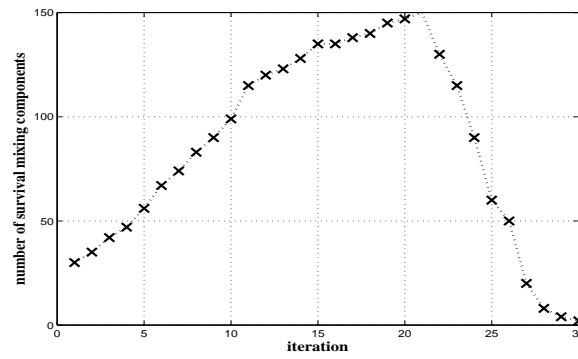


Fig. 7: The number of survival mixing components in the IS pdf in an example run of the proposed PE-SMC algorithm in searching optimum for a 20D Rastrigin function defined in Sec.VII-I

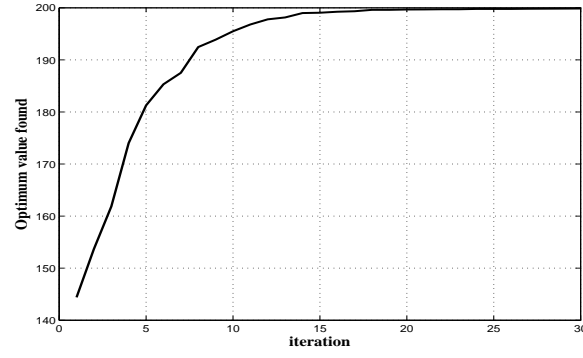


Fig. 8: The value change of the optimum found as far in an example run of the proposed PE-SMC algorithm in searching optimum for a 20D Rastrigin function defined in Sec.VII-I

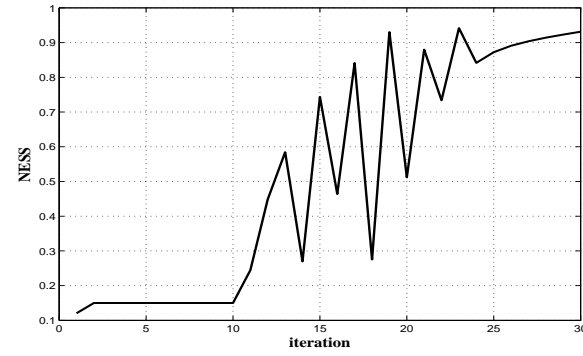


Fig. 9: The normalized ESS (NESS) value in an example run of the PE-SMC algorithm with respect to the 20D Rastrigin test function defined in Sec.VII-I

## VI. CONCLUDING REMARKS

A novel stochastic simulation based global optimization algorithm is proposed, which falls within the SMC framework. The key element of this method, which discriminates itself from the other existing alternatives, is an operator termed posterior exploration. This operator makes full use of the intermediate information yielded by the sequential importance sampling process, facilitating the exploration process to search potential promising regions in both macroscopic and microscopic scales. This

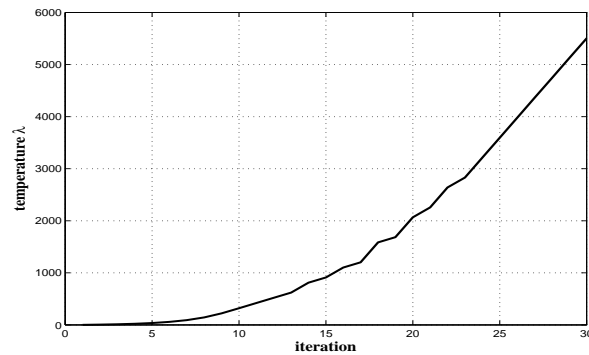


Fig. 10: The yielded annealing temperature  $\lambda$  of an example run of the PE-SMC algorithm with respect to the 20D Rastrigin test function

algorithm also provides a practical solution to tradeoff the annealing temperature change rate and the sample size. Specifically, given a preset sample size, an adaptive method based on the IS efficiency measure ESS is proposed to determine the cooling schedule online. The algorithm was applied to over a dozen benchmark test functions and compared with existing alternatives in performance. The results show that our algorithm performs satisfactorily and consistently in all cases. In particular, it outperforms the PFO and SMC-SA methods significantly and is preferable to the PSO algorithm for higher dimensional cases.

Despite being convergent in theory, the PE-SMC algorithm still has limitations, which need further investigations in the future. First, the sequence of annealed target pdfs converge on the set of global optima as the annealing variable  $\lambda$  tends to infinity, while, in practice, the algorithm will end provided that it has no better solutions output for several successive iterations. Therefore, the sampling process usually stops at a relatively big  $\lambda$  value (with respect to the initialized value for  $\lambda$ ), other than infinity. How to set an appropriate initial  $\lambda$  value also deserves further considerations. An empirical guideline is that, if the objective function  $f(\cdot)$  is essentially plateaus with a barely perceptible central peak, an initial  $\lambda \gg 1$  is preferable; otherwise, if  $f(\cdot)$  is much peaky, then an initial  $\lambda \ll 1$  is better for use. Second, as an empirical measure to evaluate the IS efficiency, the ESS may become misleading under certain circumstances, e.g., when all the samples get trapped into a local region around a local optimum that is completely isolated from the true global optima. Third, despite the great flexibility, the student's  $t$  mixture pdf is not universally adaptable to all cases. For example, if the global optima happen to lie at the boundary of the solution space, asymmetric or truncated symmetric pdfs will be better choices for constructing the IS pdf in concept. Last but not least, the current PE-SMC algorithm still suffers from curse of dimensionality. How to adapt it to deal with ultra high dimensional cases with acceptable computational burdens is still an open problem.

## VII. APPENDIX: TEST FUNCTIONS

The original forms of these test functions are devised to find minimums. We translate them correspondingly, making them suitable for testing the algorithms that are devised to find the maximum. Assume that the original function  $g(x)$  defined on  $\chi$  is upper bounded by  $g_u$ , i.e.,  $g(x) \leq g_u, \forall x \in \chi$ , then we just select a  $g_s$  that is bigger than  $g_u$  and translate the test function to be  $f(x) = g_s - g(x)$ . The modified test function is presented as follows.

### A. TF1: Ackley function

This function is defined to be

$$f(x) = 30 - \left[ -a \exp \left( -b \sqrt{\frac{1}{d} \sum_{i=1}^d x_i^2} \right) - \exp \left( \frac{1}{d} \sum_{i=1}^d \cos(cx_i) \right) + a + \exp(1) \right]. \quad (23)$$

It is characterized by a nearly flat outer region, and a large peak at the centre, thus it poses a risk for optimization algorithms to be trapped in one of its many local maxima. The global maximum is  $f(x^*) = 30$ , at  $x^* = (0, \dots, 0)$ .

### B. TF2: Cross-in-tray function

This function is defined to be

$$f(x) = -0.5 - \left( -0.0001 (|g(x_1, x_2)| + 1)^{0.1} \right), \quad (24)$$

where

$$g(x_1, x_2) = \sin(x_1) \sin(x_2) \exp \left( \left| 100 - \frac{\sqrt{x_1^2 + x_2^2}}{\pi} \right| \right).$$

It is evaluated on the square  $x_i \in [-10, 10]$ , for  $i = 1, 2$ . The global maximum  $f(\mathbf{x}^*) = 1.56261$  is located at  $\mathbf{x}^* = (1.3491, -1.3491), (1.3491, 1.3491), (-1.3491, 1.3491)$  and  $(-1.3491, -1.3491)$ .

#### C. TF3: Drop-wave function

This is multimodal and highly complex function. It is evaluated on the square  $x_i \in [-5.12, 5.12]$ , for  $i = 1, 2$ , as follows

$$f(\mathbf{x}) = \frac{1 + \cos(12\sqrt{x_1^2 + x_2^2})}{0.5(x_1^2 + x_2^2) + 2}. \quad (25)$$

The global maximum  $f(\mathbf{x}^*) = 1$  is located at  $\mathbf{x}^* = (0, 0)$ .

#### D. TF4: Eggholder function

This is a difficult function to optimize, because of a large number of local minima. It is evaluated on the square  $x_i \in [-512, 512]$ , for  $i = 1, 2$ , as follows

$$f(\mathbf{x}) = 1500 - \left[ -(x_2 + 47) \sin \left( \sqrt{\left| x_2 + \frac{x_1}{2} + 47 \right|} \right) - x_1 \sin \left( \sqrt{|x_1 - (x_2 + 47)|} \right) \right]. \quad (26)$$

The global maximum  $f(\mathbf{x}^*) = 2459.6407$  is located at  $\mathbf{x}^* = (512, 404.2319)$ .

#### E. TF5: Griewank function

This function has many widespread local minima, which are regularly distributed. It is usually evaluated on the hypercube  $x_i \in [-600, 600]$ , for all  $i = 1, \dots, d$ , as follows

$$f(\mathbf{x}) = 1000 - \left( \sum_{i=1}^d \frac{x_i^2}{4000} - \prod_{i=1}^d \cos \left( \frac{x_i}{\sqrt{i}} \right) + 1 \right). \quad (27)$$

The global maximum  $f(\mathbf{x}^*) = 1000$  is located at  $\mathbf{x}^* = (0, \dots, 0)$ .

#### F. TF6: Holder table function

This function has many local maximum, with four global one. It is evaluated on the square  $x_i \in [-10, 10]$ , for  $i = 1, 2$ , as follows

$$f(\mathbf{x}) = \left| \sin(x_1) \cos(x_2) \exp \left( \left| 100 - \frac{\sqrt{x_1^2 + x_2^2}}{\pi} \right| \right) \right|. \quad (28)$$

Its global maximum  $f(\mathbf{x}^*) = 19.2085$  is located at  $\mathbf{x}^* = (8.05502, 9.66459), (8.05502, -9.66459), (-8.05502, 9.66459)$  and  $(-8.05502, -9.66459)$ .



### G. TF7: Levy function

This function is evaluated on the hypercube  $x_i \in [-10, 10]$ , for all  $i = 1, \dots, d$ , as follows

$$f(\mathbf{x}) = 100 - \left( \sin^2(\pi w_1) + \sum_{i=1}^{d-1} (w_i - 1)^2 [1 + 10 \sin^2(\pi w_i + 1)] + (w_d - 1)^2 [1 + \sin^2(2\pi w_d)] \right). \quad (29)$$

where  $w_i = 1 + \frac{x_i - 1}{4}$ , for all  $i = 1, \dots, d$ . The global maximum  $f(\mathbf{x}^*) = 100$  is located at  $\mathbf{x}^* = (1, \dots, 1)$ .

### H. TF8: Levy function N.13

This function is evaluated on the hypercube  $x_i \in [-10, 10]$ , for  $i = 1, 2$ , as below

$$f(\mathbf{x}) = 450 - (\sin^2(3\pi x_1) + (x_1 - 1)^2 [1 + \sin^2(3\pi x_2)] + (x_2 - 1)^2 [1 + \sin^2(2\pi x_2)]), \quad (30)$$

with global maximum  $f(\mathbf{x}^*) = 450$ , at  $\mathbf{x}^* = (1, \dots, 1)$ .

### I. TF9: Rastrigin function

This function is evaluated on the hypercube  $x_i \in [-5.12, 5.12]$ , for all  $i = 1, \dots, d$ , with the form

$$f(\mathbf{x}) = 200 - \left( 10d + \sum_{i=1}^d [x_i^2 - 10 \cos(2\pi x_i)] \right). \quad (31)$$

Its global maximum  $f(\mathbf{x}^*) = 200$  is located at  $\mathbf{x}^* = (0, \dots, 0)$ . This function is highly multimodal, but locations of the local minima are regularly distributed.

### J. TF10: The second Schaffer function

This function is usually evaluated on the square  $x_i \in [-100, 100]$ , for all  $i = 1, 2$ . It has a form as follows

$$f(\mathbf{x}) = 1 - \left( 0.5 + \frac{\sin^2(x_1^2 - x_2^2) - 0.5}{[1 + 0.001(x_1^2 + x_2^2)]^2} \right). \quad (32)$$

Its global maximum  $f(\mathbf{x}^*) = 1$  is located at  $\mathbf{x}^* = (0, 0)$ .

### K. TF11: Schwefel function

This function is also complex, with many local minima. It is evaluated on the hypercube  $x_i \in [-500, 500]$ , for all  $i = 1, \dots, d$ , as follows

$$f(\mathbf{x}) = 1800 - \left( 418.9829d - \sum_{i=1}^d x_i \sin(\sqrt{|x_i|}) \right). \quad (33)$$

Its global maximum  $f(\mathbf{x}^*) = 1800$  is located at  $\mathbf{x}^* = (420.9687, \dots, 420.9687)$ .

### L. TF12: Shubert function

This function has several local minima and many global minima. It is usually evaluated on the square  $x_i \in [-10, 10]$ , for all  $i = 1, 2$ .

$$f(\mathbf{x}) = 300 - \left( \sum_{i=1}^5 i \cos((i+1)x_1 + i) \right) \left( \sum_{i=1}^5 i \cos((i+1)x_2 + i) \right). \quad (34)$$

Its global maximum is  $f(\mathbf{x}^*) = 486.7309$ .

*M. TF13: Perm Function  $0, d, \beta$*

This function is evaluated on the hypercube  $x_i \in [-d, d]$ , for all  $i = 1, \dots, d$ , as follows

$$f(\mathbf{x}) = 120 - \left( \sum_{i=1}^d \left( \sum_{j=1}^d (j + \beta) \left( x_j^i - \frac{1}{j^i} \right) \right)^2 \right). \quad (35)$$

Its global maximum  $f(\mathbf{x}^*) = 120$  is located at  $\mathbf{x}^* = (1, \frac{1}{2}, \dots, \frac{1}{d})$ .

*N. TF14: Rosenbrock function*

This function is also referred to as the Valley or Banana function. The global maximum lies in a narrow, parabolic spike. However, even though this spike is easy to find, convergence to the maximum is difficult [33]. It has the following form

$$f(\mathbf{x}) = 1.8e5 - \sum_{i=1}^{d-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2], \quad (36)$$

and is evaluated on the hypercube  $x_i \in [-5, 10]$ , for all  $i = 1, \dots, d$ . The global maximum  $f(\mathbf{x}^*) = 1.8e5$  is located at  $\mathbf{x}^* = (1, \dots, 1)$ .

*O. TF15: The fifth function of De Jong*

The fifth function of De Jong is multimodal, with very sharp drops on a mainly flat surface. The function is evaluated on the square  $x_i \in [-65.536, 65.536]$ , for all  $i = 1, 2$ , as follows

$$f(\mathbf{x}) = 510 - \left( 0.002 + \sum_{i=1}^2 5 \frac{1}{i + (x_1 - a_{1i})^6 + (x_2 - a_{2i})^6} \right)^{-1}, \quad (37)$$

where

$$\mathbf{a} = \begin{pmatrix} -32 & -16 & 0 & 16 & 32 & -32 & 9 & 16 & 32 \\ -32 & -32 & -32 & -32 & -32 & -16 & 32 & 32 & 32 \end{pmatrix}.$$

The global maximum  $f(\mathbf{x}^*) \approx 509$  [31].

*P. TF16: Easom function*

This function has several local maximum. It is unimodal, and the global maximum has a small area relative to the search space. It is usually evaluated on the square  $x_i \in [-100, 100]$ , for all  $i = 1, 2$ , as follows

$$f(\mathbf{x}) = 0 - \cos(x_1) \cos(x_2) \exp \left( -(x_1 - \pi)^2 - (x_2 - \pi)^2 \right). \quad (38)$$

The global maximum  $f(\mathbf{x}^*) = 1$  is located at  $\mathbf{x}^* = (\pi, \pi)$ .

### Q. TF17: Michalewicz function

This function has  $d!$  local minima, and it is multimodal. The parameter  $m$  defines the steepness of the valleys and ridges; a larger  $m$  leads to a more difficult search. The recommended value of  $m$  is  $m = 10$ . It is usually evaluated on the hypercube  $x_i \in [0, \pi]$ , for all  $i = 1, \dots, d$ , as follows

$$f(\mathbf{x}) = - \sum_{i=1}^d \sin(x_i) \sin^{2m} \left( \frac{ix_i^2}{\pi} \right). \quad (39)$$

The global maximum in 2D case is  $f(\mathbf{x}^*) = 1.8013$  located at  $\mathbf{x}^* = (2.20, 1.57)$ , in 5D case is  $f(\mathbf{x}^*) = 4.687658$  and in 10D case is  $f(\mathbf{x}^*) = 9.66015$ .

### REFERENCES

- [1] J. Kennedy, "Particle swarm optimization," in *Encyclopedia of Machine Learning*. Springer, 2010, pp. 760–766.
- [2] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm intelligence*, vol. 1, no. 1, pp. 33–57, 2007.
- [3] D. Whitley, "A genetic algorithm tutorial," *Statistics and computing*, vol. 4, no. 2, pp. 65–85, 1994.
- [4] M. Dorigo and M. Birattari, "Ant colony optimization," in *Encyclopedia of machine learning*. Springer, 2010, pp. 36–39.
- [5] M. Dorigo, M. Birattari, and T. Stützle, "Ant colony optimization," *Computational Intelligence Magazine, IEEE*, vol. 1, no. 4, pp. 28–39, 2006.
- [6] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi *et al.*, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.
- [7] L. Ingber, "Simulated annealing: Practice versus theory," *Mathematical and computer modelling*, vol. 18, no. 11, pp. 29–57, 1993.
- [8] E. Zhou and X. Chen, "Sequential monte carlo simulated annealing," *Journal of Global Optimization*, vol. 55, no. 1, pp. 101–124, 2013.
- [9] M.-S. Oh and J. O. Berger, "Adaptive importance sampling in monte carlo integration," *Journal of Statistical Computation and Simulation*, vol. 41, no. 3–4, pp. 143–168, 1992.
- [10] O. Cappé, R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert, "Adaptive importance sampling in general mixture classes," *Statistics and Computing*, vol. 18, no. 4, pp. 447–459, 2008.
- [11] R. M. Neal, "Annealed importance sampling," *Statistics and Computing*, vol. 11, no. 2, pp. 125–139, 2001.
- [12] P. Del Moral, A. Doucet, and A. Jasra, "Sequential monte carlo samplers," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 3, pp. 411–436, 2006.
- [13] M. Hauschild and M. Pelikan, "An introduction and survey of estimation of distribution algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 3, pp. 111–128, 2011.
- [14] A. Auger and N. Hansen, "Tutorial CMA-ES: evolution strategies and covariance matrix adaptation," in *GECCO*, 2012, pp. 827–848.
- [15] O. Krause and C. Igel, "A more efficient rank-one covariance matrix update for evolution strategies," in *Proc. of the 2015 ACM Conf. on Foundations of Genetic Algorithms XIII*. ACM, 2015, pp. 129–136.
- [16] H.-G. Beyer and S. Finck, "On the design of constraint covariance matrix self-adaptation evolution strategies including a cardinality constraint," *Evolutionary Computation, IEEE Transactions on*, vol. 16, no. 4, pp. 578–596, 2012.
- [17] R. Y. Rubinstein and D. P. Kroese, *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media, 2013.
- [18] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.
- [19] E. Zhou, M. C. Fu, S. Marcus *et al.*, "Particle filtering framework for a class of randomized optimization algorithms," *Automatic Control, IEEE Transactions on*, vol. 59, no. 4, pp. 1025–1030, 2014.
- [20] X. Chen and E. Zhou, "Population model-based optimization with sequential monte carlo," in *Proc. of the 2013 Winter Simulation Conference*. IEEE Press, 2013, pp. 1004–1015.
- [21] B. Liu, S. Cheng, and Y. Shi, "Particle filter optimization: A brief introduction," in *Advances in Swarm Intelligence*, 2016, pp. 95–104.
- [22] B. Liu, "Adaptive annealed importance sampling for multimodal posterior exploration and model selection with application to extrasolar planet detection," *The Astrophysical Journal Supplement Series*, vol. 213, no. 1, p. 14, 2014.

- [23] H. Haario, E. Saksman, and J. Tamminen, "Componentwise adaptation for high dimensional mcmc," *Computational Statistics*, vol. 20, no. 2, pp. 265–273, 2005.
- [24] J. Geweke, "Bayesian inference in econometric models using monte carlo integration," *Econometrica: Journal of the Econometric Society*, pp. 1317–1339, 1989.
- [25] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.
- [26] G. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2004.
- [27] J. S. Liu and R. Chen, "Sequential monte carlo methods for dynamic systems," *Journal of the American statistical association*, vol. 93, no. 443, pp. 1032–1044, 1998.
- [28] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of nonlinear filtering*, vol. 12, no. 656-704, p. 3, 2009.
- [29] I. C. Trelea, "The particle swarm optimization algorithm: convergence analysis and parameter selection," *Information processing letters*, vol. 85, no. 6, pp. 317–325, 2003.
- [30] K. E. Parsopoulos and M. N. Vrahatis, "Parameter selection and adaptation in unified particle swarm optimization," *Mathematical and Computer Modelling*, vol. 46, no. 1, pp. 198–213, 2007.
- [31] X. Yao, Y. Liu, and G. Lin, "Evolutionary programming made faster," *Evolutionary Computation, IEEE Transactions on*, vol. 3, no. 2, pp. 82–102, 1999.
- [32] B. Hajek, "Cooling schedules for optimal annealing," *Mathematics of operations research*, vol. 13, no. 2, pp. 311–329, 1988.
- [33] V. Picheny, T. Wanger, and D. Ginsbourger, "A benchmark of kriging-based infill criteria for noisy," *Struct Multidiscip Optim*, vol. 10, p. 1007, 2013.